Assignment II : Machine Learning

July 14, 2025

ID5003W: Industrial AI at Scale Laboratory

Total Marks:50

Submission Deadline : July 22

Objective : Use the Bank Marketing dataset (saved in drive) to build a binary classifier that predicts whether a customer will subscribe to a term deposit using Spark MLlib.
Do the following tasks on the dataset:

Problem 1 : Data Understanding (Marks : 10)

1. Load the dataset and print schema
2. Count how many customers subscribed (y = yes) vs not.
3. Print the distinct values of job and education and count of each value.
4. split the data in train and test dataset and report distribution of output values in train and test dataset
5. Use your favorite strategy to balance the dataset, if you find that data is imbalanced and is important to balance the dataset before training the model. Provide explanation of your decision.

Problem 2: Data Preprocessing (Marks : 10)

1. Identify categorical and numerical columns.
2. Use StringIndexer and OneHotEncoder on all categorical features.
3. Assemble all features using VectorAssembler.
4. Use StringIndexer on the label column y.

Problem 3: Model Building (Marks : 10)

1. Build a pipeline using Logistic Regression.
2. Train a random forest.
3. Evaluate using accuracy, precision, and recall.
4. Print the name of the best model and show confusion matrix for the best model.

Problem 4: Hyperparameter tuning and parallelism (Marks : 10)

1. Select hyperparameters for tuning and their values and justify your selection
2. Perform cross validation for random forest.
3. Compare the accuracy results of logistic regression, random forest without hyperparameter tuning, random forest with hyperparameter tuning models.

Problem 5: Performance Profiling (Marks : 10)

1. For each of the following configurations, measure the run time taken by each section and display the results in a table:
   (a) `spark.executor.cores = 1`, `spark.max.cores = 2`, `spark.executor.memory = 1g`
   (b) `spark.executor.cores = 2`, `spark.max.cores = 2`, `spark.executor.memory = 1g`
   (c) `spark.executor.cores = 2`, `spark.max.cores = 2`, `spark.executor.memory = 1g`, with cross-validation parallelism = {1, 2}
2. Write your observations on the run times and explain the effect of parallelism on performance. How does increasing parallelism impact the runtime?

Execution Instructions

1. Use the data provided in the folder to implement the problems on google colab
2. Change the file path before submitting the job to server to run on spark cluster
3. Download the file as '.py' file
4. configuration of spark.executor.cores and spark.cores.max is must in the code to run it on the lab cluster. The value of spark.cores.max should be divisible by spark.executor.cores (i.e. spark.executor.instances = spark.cores.max/spark.executor.cores)
5. execution command for server : python ./spark_job_execution.py ./your_code.py your_rollno
6. Check the log file for the output

Submission Instructions

1. Submit the log files.
2. Submit a report with details on the output generated for each problem, and name of logfiles. Notebook is not accepted.
3. Submit the code that you have submitted to spark cluster.